DOCUMENT RESUME

ED 108 218 CS 202 103

TITLE	Common Sense and Testing in English; Report of the Task Force on Measurement and Evaluation in the Study of English.
INSTITUTION	National Council of Teachers of English, Urbana, Ill.
PUB DATE	75
NOTE	40p.
AVAILABLE FROM	National Council of Teachers of English, 1111 Kenyon Rd., Urbana, Illinois 61801 (Stock No. 07737, $1.00)
EDRS PRICE	MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS	Elementary Secondary Education; *English Instruction; Evaluation; Measurement Instruments; Norm Referenced Tests; Standardized Tests; *Testing; *Test Reliability; *Tests; *Test Validity

ABSTRACT

This report of the Task Force on Measurement and Evaluation in the Study of English (appointed by the National Council of Teachers of English) analyzes the present state of the art of testing and recommends the use of common sense in selecting and using tests and in interpreting the information derived from testing. The report views standardized tests in current use as inadequate and advises makers of tests to be more explicit in stating the limitations of their tests, to do more to describe the populations on which norms are based, and to be more effective in informing users about proper and improper uses of tests and test data. Among the alternatives which the Task Force suggests as superior to testing are portfolios of student work, interviews, peer evaluation, and observations of classroom performance. (JM)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED108218

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Common Sense and Testing in English

Report of the Task Force on Measurement and Evaluation in the Study of English

Alan Purves, University of Illinois, Chairman
Beryl Bailey, City University of New York
Robert A. Bennett, San Diego Unified Schools
Charles W. Daves, Educational Testing Service
Helen Lodge, California State University, Northridge
Olive S. Niles, State Department of Education, Hartford, Connecticut
Roy C. O'Donnell, University of Georgia
Leo P. Ruth, University of California, Berkeley
Richard L. Venezky, University of Wisconsin, Madison
Robert E. Beck, John Swett High School, Crockett, California, *ex officio*

National Council of Teachers of English
1111 Kenyon Road
Urbana, Illinois 61801

CS202103

EDITORIAL BOARD Charles R. Cooper, Evelyn Copeland, Bernice E. Cullinan, Richard Lloyd-Jones, Frank Zidonis, Robert F. Hogan, *ex officio,* Paul O'Dea, *ex officio*

COVER DESIGN Bob Bingenheimer
STAFF EDITOR Carol Schanche
STAFF TYPESETTER Carol J. Shore

NCTE Stock Number 01116

Copyright © 1975 by the National Council of Teachers of English.
All rights reserved. Printed in the United States of America.

National Council of Teachers of English
1111 Kenyon Road, Urbana, Illinois 61801

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
National Council of Teachers of English
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

CONTENTS

iv	Introduction
1	The Facets of English
3	Uses and Misuses of Test Information
7	Measurement of the Outcomes and Processes of Instruction in English
12	Standardized Tests and the Measurement of English Instruction
17	Criterion-Referenced and Domain-Referenced Tests and the Measurement of English Instruction
19	Criteria for Selecting Standardized Tests in English
26	Criteria for the Interpretation and Use of Test Results
29	Appendix A. Checklist for Evaluating English Tests and Test Uses
31	Appendix B. Citizen's Edition: Common Sense and Testing in English

INTRODUCTION

Teachers of English and language arts have always wanted to know the results of their teaching, to see if their efforts have done someone some good. One of the devices they have used for gauging progress has been the standardized test.

The relation of the teacher of English to the standardized test, however, has been an uneasy alliance. The teacher has known, for example, that standardized English tests do not measure what test titles often imply. They have known that there simply is no one "English" or language arts test, because English is a subject of many parts, and only many kinds of tests could begin to measure student progress in English. Tests of spelling, tests of usage, tests of knowledge or grammar, tests of literary history, tests of knowledge about rhetoric and a dozen other topics can provide only a beginning of knowing how students have grown.

The teacher of English language arts has placed far greater reliance on the fabled gleam in a student's eye, the subtle signs that Tom in the front row, because of a story he has read, has begun to grasp the complexity of relationships between two people he knows, and that Amanda has finally begun to learn how to say on paper what she is so capable of saying in spoken language. There are no standardized tests for what Tom and Amanda have learned.

The teacher of English language arts has known that standardized tests, with their appurtenances of statistics, precision of questioning, and norming procedures can help with knowing how students are progressing and has welcomed that help. But recent trends in education—the current call for "accountability" for example—have begun to raise the standardized test to vital importance, and a great deal of nonsense about standardized tests and their use has consequently been let loose on the land. The members of the National Council of Teachers of English believe it is imperative that the public and the profession know that great harm is being done to students and to their education by unwarranted faith in standardized tests.

A resolution to this effect was passed at the Annual Business Meeting of NCTE in Philadelphia in November 1973. With the help of the NCTE Research Foundation, a blue-ribbon Task Force was called in the summer of 1974, and this booklet is the result of their work.

At one point, the booklet was called "Truth in Testing," a title which evoked images of Ralph Nader and other consumer advocates challenging the kinds of fraud which unscrupulous manufacturers visit upon consumers. But further study led the writers of this report to abandon that image because the testing fraud is in major part something that is done by the consumers to themselves. The investigation by the Task Force on Measurement and Evaluation in the Study of English revealed that most makers of tests give reasonable warning to users that their tests are based on limited samplings of students, that the norms give, at best, only approximate estimates of student performance, and that the tests are very limited in scope and cannot correspond with local curriculum. When such warnings are given to users, the test-maker cannot be held completely liable for the misuses of the data.

The Task Force found evidence of widespread ignorance about tests among teachers, administrators, members of school boards, the media, and the public.

From such ignorance only folly can flow and the only antidote to that ignorance is knowledge about the present state of the art of testing. That is what this booklet is about.

The Task Force report does not intend a condemnation of standardized tests, but it does intend that people who use tests know the fragile (and sometimes corrupt) information that standardized tests supply. The report calls for using common sense in selecting and using tests and interpreting the information derived from testing. At the same time, it calls for the makers of tests to be more explicit in stating the limitations of their tests, to do more to describe the populations on which norms are based, and to be more effective in informing users on proper and improper uses of tests and test data.

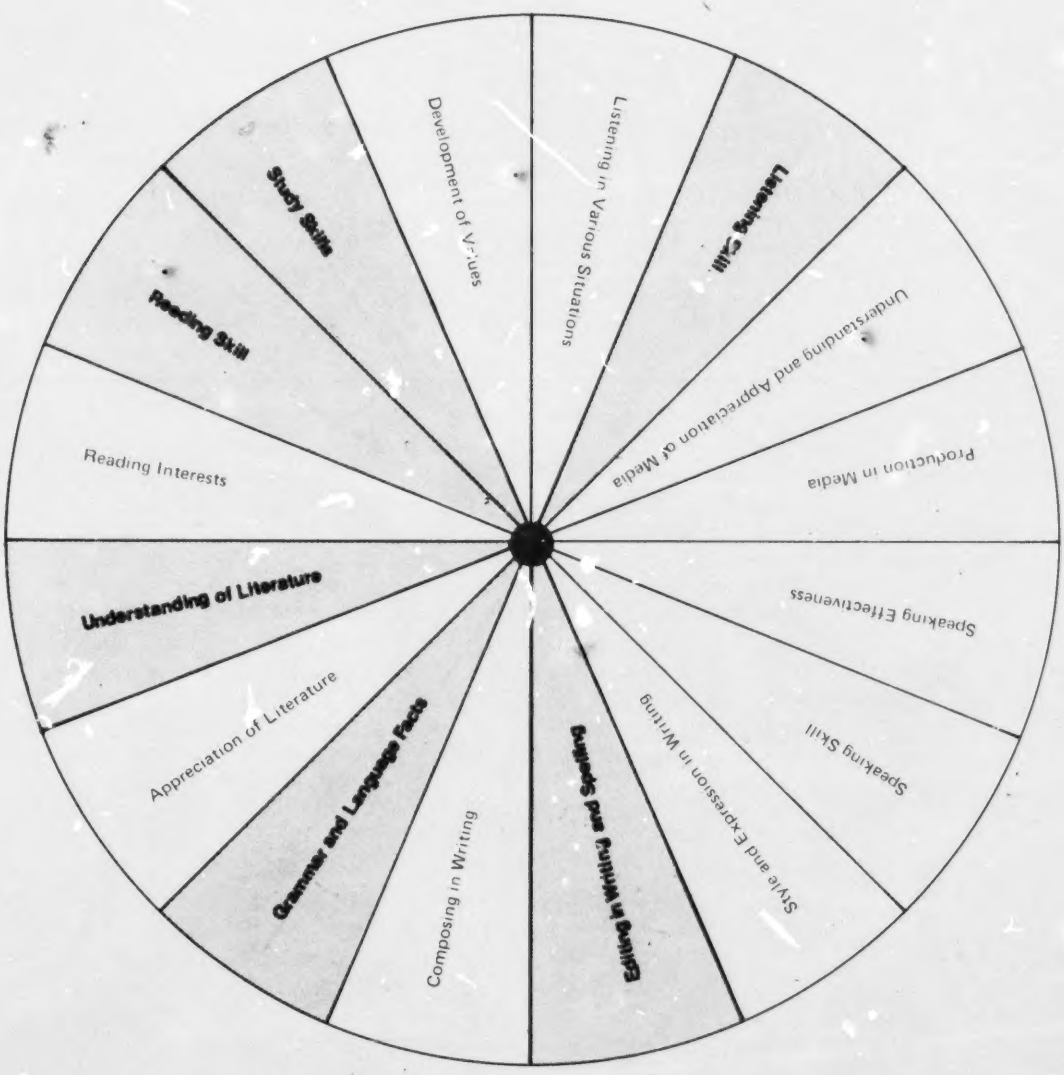
The report is also a directive and an appeal to the members of the teaching profession to become part of the means by which common sense about testing is put to work in the schools and the society. The American fascination with numbers is a dangerous tendency which, for example, allows us to drift into belief that the magic number "7.8" tells us that the student making that score in a standardized reading test "belongs" in the eighth month of the seventh grade. It does nothing of the kind. The "grade-level" myth is a harmful flowering of our faith in numbers, and is particularly attacked by the Task Force.

Development of a public disposition to use common sense in our contacts with testing can halt the insidious trend to use standardized tests as main instruments for judging educational programs. The "Citizen's Edition" in the appendix of this booklet is given to the public and the profession as one device for building widespread understanding of the limits of standardized tests. Teachers and other citizens are invited to prepare copies of this brief statement for parent-teacher meetings, for inservice programs, for informing the media, for telling students the truth about the tests they take. But this booklet and the appended citizen's edition will lie unknown and unused unless you and others like you do something about it.

While this booklet is about tests in English language arts, its analysis of the art of testing has insights for all subjects and all subject-matter teachers. The National Council of Teachers of English believes that problems with uses of standardized tests are particularly acute for our subject, but we know that our problems with testing are not unique. Therefore, this booklet is offered for the benefit of the profession at large, for the students whose lives can be changed by uses of standardized tests, and for the citizens who pay for schools and must judge what we do.

I urge associations of English language arts teachers, departments and faculties of elementary and secondary schools, school district evaluation committees, and other concerned groups to take up the information in this booklet and begin a searching public dialogue about the uses and abuses of standardized tests. That dialogue should begin now.

John C. Maxwell
Deputy Executive Secretary
National Council of Teachers of English



Some of the Many Facets of English

• The shaded sections are those partially covered by standardized English Tests •

THE FACETS OF ENGLISH

Evaluating student performance in English is complex and requires some identification of the areas that comprise the discipline we call English. Certainly English deals with the variety of ways people use their native or second language to communicate with each other, to express themselves, or to create art. English also deals with the ways people have of describing these uses of language; it deals with the connection of language and the psychology and culture of the person or persons using it. English includes the examination of the heightened use of language in literature, which combines the expression of an author's personality and culture with universals of human experience. English includes skills, information, attitudes, and feelings that people possess about language and about those who have used or now use language. English deals with varieties of written and oral language, as well as with the related nonverbal modes of communication and expression, and with varieties of media. Finally, English deals with people's growth in language ability and growth through language to self-awareness and self-assurance.

What Is Taught

Many of these facets of English are the explicit concerns of schools; many are not. Many are the general concerns of all teachers in school; many are the specific concerns of English teachers. These specific concerns have usually included reading, writing, speaking, listening, study skills, literature, the use of media, information about language and the people who have used it, and the relationship of an individual's language to the language of the various cultures around that individual. Teachers have tried to help students acquire and develop skills, to weigh concepts and values and to nurture attitudes, interests, and habits of behavior.

What Is Tested

Of all this variety, only a small part has been the focus of testing—an even smaller part the focus of "standardized" testing. Even so, tests have a way of determining educational priorities, and the increased use of standardized tests has regrettably shrunk the list of priorities available for people to consider. Standardized tests* have traditionally dealt with the more easily measurable aspects of English instruction, to the neglect of the full range of activities—involving thoughts, feelings, and attitudes—that are stressed by English teachers. This situation was acceptable, perhaps, as long as educators recognized these limitations and as long as standardized testing was not relied upon in making far-reaching decisions in educational policy. The demand for accountability in

*_Standardized Test:_ A test in which all students answer a number of questions under the same directions and time conditions. The scores of one group on the test may then be compared to the scores of another group. It should not be inferred that a standardized test sets standards for achievement.

education, however, has drastically changed the role of measurement and made it more and more central in the management of education. At present, standardized tests do not provide a valid or appropriate basis for many decisions. They cannot provide the necessary information to resolve such major issues as modification of the content of the English curriculum, changes in methods of teaching, and evaluation of the competence of teachers. It is doubtful whether they can ever do so, even assuming the possibility of great improvements in the tests themselves.

USES AND MISUSES OF TEST INFORMATION

Many individuals and groups have quite legitimate interests in how students are progressing and what kind of job the schools are doing. They want to know that student progress is substantial and sustained. Recently, the term "accountability" has come into use to describe these interests and the responsibility of the schools to provide progress reports on the effect of instruction in the schools. But not everybody wants the same kind of information, and it helps to sort out the legitimate uses and users of test results.

Students. At nearly every point during their schooling, some students want to know, and have a right to know, how they are doing. Are they getting better at what they are supposed to be getting better at? Are they meeting the teacher's expectations? How are they doing in comparison to their class or group? Are they becoming qualified for some important next step in their schooling? Test scores give limited answers to these questions, but answers that can have an important effect on a student's self-concept as well as academic success. Some students are not interested in their progress; some can be unhealthily competitive. Teachers must recognize real concerns without encouraging harmful competition, and must realize that often students' uses of test data are legitimate.

Parents. Many parents ask the same questions that their children ask. Of course they are entitled to know how their children are doing. As long as school is seen as a means of advancement or a way of securing a place in society and as long as parents have hopes and fears for their children, such questions will arise. Achievement as indicated by some form of test scores can provide a partial answer to these questions—but by no means a complete answer.

Teachers. A teacher may have several uses for test information. One comes from concern about the success of teaching. Have the students met the objectives of the course of study? What appear to be the points that demand additional instruction? A second use comes from a concern with a specific student's success or failure so that additional or remedial instruction may be carried out. At times teachers may want to compare approaches; test results can play only a modest part in this comparison, and need to be supplemented with other kinds of information, such as details on interests and attitudes.

School Administrators. Principals, department heads, school psychologists, or guidance counselors may use test results as one source of information to

---

**Measurement and Grading:* Two terms that are frequently used and sometimes confused. *Measurement* is the procedure by which human behavior or human characteristics are described, usually in mathematical terms. Measurement depends upon some formal or informal way of recording information about a person (such as getting the person to write a paper or take a multiple-choice test) and upon some way of scoring or rendering that information (as in number of errors or percentage right). *Grading* is a term given to the attempt by teachers to sum up their estimates of students and their work. Grades are given either relating the student's work to previous work, to some ideal, or to other students.

help determine the success of individual students, of classes, and of the school as a whole. Decisions based in part on test results may have to do with providing supplementary instruction for one student or advancing another to a higher level. Other decisions may have to do with evaluating* a particular instructional program or with providing remedial or alternate instruction for some number of students.

District and Community. A community, particularly through its board of education, may be concerned with the results of tests as they provide information to help decide how well individual schools or the system as a whole is meeting the specified educational needs of the community. Other uses may be to allocate money, people, or effort and to evaluate alternatives in curriculum, organization, and instruction.

States. A state government is concerned with the effectiveness and fairness of its educational programs. In a few instances, a state has used test results to help decide which types of communities need special support. It may also try to use them as part of an assessment* of needs in education, and of the success of the schools in satisfying those needs.

Institutions of Higher Education. Colleges and universities use school test results in two different ways. The most common is to provide evidence of rank in test performance to help determine admission to and, in some cases, placement in a particular group within the institution. A second use is to provide data for research in curriculum and instruction in the elementary and secondary schools, for the evaluation of new concepts and methods in comparison with older ones.

Publishers and Other Producers of Educational Materials. In rare instances this large group will use test results to determine the effectiveness of their materials or to determine what new ventures to undertake. If, for example, certain groups of students are not doing well on certain tests, an organization may use this information as a reason for developing instructional materials for those groups.

**Evaluation:* The process of determining the value or worth of something. In schooling at any level, evaluation may be of students or of instructional programs. Often the same scores are used to test students' performance and attitudes and to evaluate programs. There are dangers in this process, because it oversimplifies the nature of instructional programs which may be concerned with more than test results.

**Assessment:* A variety of measurement, now a term used for local, state, and national projects that seek to describe how well students are doing in various fields. Assessment reporting may be likened to reporting from polls: both report in percentages. Some assessment programs do not yield individual scores, but information about competence (e.g., the percentage of twelfth graders that can write a clear set of directions). Assessment projects are not designed to provide scores for individual students or individual schools. The National Assessment of Educational Progress is the largest ongoing assessment program.

The Federal Government. The interest of the Federal Government in test results stems from a mandate to insure fairness of educational opportunity to all citizens. Test results may help to locate inequities. They might conceivably serve as indicators of the success of federally sponsored projects or the need for modification of those projects in the light of local conditions.

The General Public. This diverse group may use test results as part of its determination of the educational well-being of the society. Test results form part of a mosaic of information that society uses to see how it is doing. It can only see how it is doing if it knows the validity of the tests and measures that are used. To date, standardized test results form the main source of information to the general public.

LEGITIMATE USES AND USERS OF THE RESULTS OF MEASUREMENT

RESULTS ABOUT THE . . .	MAY BE USED BY THE . . .	TO HELP . . .
student	student student's parents student's teacher student's counselor	evaluate individual progress make decisions about teaching guide the student into appropriate programs
class	students in class teacher of class building administrator	evaluate class progress improve programs
school	students in school teachers in school building administrators district administrators	evaluate programs identify needs
district	teachers in district administrators in district parents and community of district state administrators	evaluate programs identify needs
state	district administrators state administrators federal officials teachers general public	identify needs assess progress
nation	state administrators federal officials teachers general public	identify needs assess progress

Misuses of Test Information

The fact that there are numerous kinds of information sought by different groups of people means that getting the wrong kind of information or using information erroneously is a common hazard. The most frequent hazard, however, is the misuse of test results. Many standardized achievement tests were designed to compare* groups of students with respect to very generally defined knowledge or skill. But the results of these tests, without other pertinent information, are commonly misused three ways: to place individual students in particular kinds of classes, to evaluate the effectiveness of a new curriculum, or to assess the strengths and weaknesses of a school system. Some other kinds of tests have been designed to indicate how far along the path to a particular goal an individual student has progressed; if the results of these tests are used to relate the achievement of one student or group of students to all the students at an age level, that, too, is a misuse of test information.

Another common misuse of test information occurs when someone takes the average test scores on a single standardized achievement test and uses them to evaluate the success of a particular teacher, school, or district. Averages are poor indicators of success; averages on tests that only partially cover what is taught are misleading indicators of success. There are much better ways to hold teachers and schools accountable.

A third misuse relates to the use of test results in admission to selective institutions, particularly colleges, universities, and professional schools. The tests used for these purposes are normally held secret to prevent "teaching for the test" with the consequence that teachers in both schools and the higher institutions cannot determine the appropriateness of the tests to their curricula. Their results, however, are released in such a way that the public thinks the tests appropriate.

---

*Comparing Students: The most common ways of comparing the scores of students are through *grade level* or *age level* scores, through *percentile* scores, or through *stanines*. *Grade level* and *age level* scores are derived by first testing a large number of students and gathering information on the age or grade of each student. Then average scores for each age or grade division are figured (e.g., students twelve years and three months average 63; students twelve years and six months average 70). If, four years later, any student scores 63, that student is reported as scoring at the twelve year and three month level. As one can see, the reasoning is somewhat circular, and assumes little change in what students learn from year to year. Preferable is *percentile* ranking, which indicates the percentage of students of a similar age or grade who scored below a given student. Thus a student score in the 91st percentile is one which is above 91% of the scores of a similar group of students. *Stanine scores* come from dividing the group of scores into nine parts. The fifth stanine is the central area where the middle 20% have scored, and the remaining stanines indicate the band of scores where decreasing percentages fall on either side of the middle point.

MEASUREMENT OF THE OUTCOMES AND PROCESSES OF INSTRUCTION IN ENGLISH

Is English an Unmeasurable Subject?

Being able to get valid and accurate instruments to measure student achievement in English depends upon one's ability to specify what is to be measured and what will be accepted as evidence that the thing to be measured has been achieved. Specifying what purpose the test is to serve also guards against inappropriate use of the results of the measurement. Measuring people's uses of their own language is not like measuring height or weight; there is necessarily some inexactness, some room for personal judgment. Indeed, much of the measurement is guess work. We estimate how a student thinks or feels from what the student says or writes, and we cannot be sure our estimate is correct.

Many of the goals of English teaching deal with subtle aspects of a person's encounters with language. Teachers want their students to be able to express their inner feelings, to be moved to tears or joy by great works of literature, to respect the language, and to develop lifelong habits and values. Although many students achieve these goals, it is not always apparent that they do so. Teachers have often sought proof and have gotten it in occasional pieces of writing or other testimonials both in school and later, but these testimonials are infrequent and teachers often have difficulty in separating the sincere from the insincere. Some teachers and researchers have sought other measures, often indirect ones, to get further proof. A larger number seek to deal only with those goals that offer more tangible proof and treat these other important goals more as desires than as readily measurable outcomes.

The thing to be measured may be a skill, a bit of knowledge, an attitude. Evidence that the thing to be measured has been achieved may be in the form of a written composition, an oral recitation, a response to a question, or some other product or behavior that can be perceived by an evaluator. The evidence may be interpreted according to a fixed scale in some instances and according to personal judgment in other instances.

If we agree, for example, that skill in punctuation, spelling, and sentence construction are desirable outcomes of instruction in English, we can either find or devise objective tests* that do a reasonably good job of measuring the extent to which our students have acquired these skills. We can usually agree on whether a student uses commas and periods in conventionally accepted ways and whether that student's spelling and sentence structure conform to accepted standards. On the other hand, ability to deal creatively with ideas and to express them in an aesthetically pleasing manner is more difficult to measure within the closed

---

*_Objective Test:_ A term that has come to be applied to those tests in which a person marks a correct answer or fills in a blank with a correct answer. The correct answer to each question has been determined in advance by a consensus of judges. The label "objective" identifies a particular kind of test format; it does not mean that the test is free of bias, error, or arbitrary decisions of correctness. Strictly speaking, there is no such thing as an objective test.

format of objective tests. We cannot use objective tests to tell the extent to which students use the resources of the English language to express themselves effectively nor the extent to which students appreciate the skillful and artistic use of language by someone else. There are other ways to determine these aspects of student achievement.

Types of Evaluation and Their Uses

Tests and other forms of measurement may be used to help to assess the instructional needs of students (called diagnostic testing), to help guide the ongoing process of instruction (called formative evaluation), or to help evaluate outcomes after instruction has been completed (called summative evaluation). Results of these measures may be used to compare one student with another student or to compare a student's performance at two different points in time. They may be used to evaluate a student in relation to a fixed standard, or to describe student interests and attitudes toward English or the school. They may also be used within a school to compare one instructional program with another or to assess the effectiveness of a program in relation to specified criteria.

What follows is a list of ways by which one may measure the growth of students individually or as groups. The information gathered can be highly accurate and highly valid, particularly when it fulfills three criteria—clarity of objectives, appropriateness of response mode, and defined criteria for judging of responses. Such information might well be presented to the various groups that have legitimate interests in indicators of student achievement, and should be received by those groups with as much respect as is now accorded the results of standardized tests. Both quantitative information from tests and qualitative information from other kinds of measures need to be combined if a true picture of student growth in English is to be determined.

Teacher-Made Tests and Departmental Tests. Whether designed for one class or all the classes in a department, teacher-made tests are good to the extent that they relate to the objectives for instruction and define what the teacher will accept as an indication of mastery of those objectives. Some teachers look to see how the average or bright students do and then adjust the grades for other students accordingly; this practice is questionable if mastery of the subject matter is the teacher's goal for all students. In their tests, teachers should emphasize program priorities and be concerned more with individual student achievement and mastery than with comparison of members of the class. In making the test, a teacher should remember that clarity, readability, and sense are as important for the quiz given to find out how students are progressing as for the test given at the end of a unit to measure instructional outcomes. In fact, teachers should be encouraged to make up their tests at the beginning of the course or unit. Doing so can help clarify objectives and their weighting. Certainly the test might well be modified during the course of instruction, but it serves as a reminder to the teacher of what the objectives are and where the instruction is going in relation to those objectives.

Grading of Papers and Other Work. A second time-honored and adequate measure of student achievement is the teacher's assessment of written assignments. Experienced English teachers develop an accurate sense of the good, the fair, and the inadequate paper. Many are able to make their criteria explicit to other teachers and to students; the more explicit those criteria are, the fairer they are for students. There are many ways of marking papers (single grades, grades split between form and content, grades according to some rating scale, and comments without grades, to mention a few) and it is not the purpose of this document to judge among them. Suffice it to say grading should be fair and not capricious, should recognize the full range of students' abilities without penalizing them for trivial lapses, and should reflect the objectives of instruction in English. Many schools successfully use teams of teachers to judge the papers of a class, thus providing opinions other than those of the classroom teacher.

Self-Assessment of Students. Often students are good judges of their own progress. A fair response to "How am I doing?" may be "How do you think you are doing?" Many students can tell what their specific strengths and weaknesses in a given area are; many can tell how much they understand of what the teacher is seeking to convey. This information is often highly accurate and may profitably be sought, particularly for diagnosing learning problems and assessing the ongoing success of instruction.

Peer Evaluation. Students can also be quite adept at judging the value or acceptability of the work of other students. Particularly in those aspects of English in which overt performance is demanded—writing, formal speaking, drama, and the like—students can be highly critical audiences. Students, of course, need both training and encouragement to help them develop and understand the nature of the criteria being applied to their own work and the work of their peers.

Questionnaires and Rating Scales That Describe Students. Teachers may undertake to measure the classroom work of students. Often they do so informally without clearly announced criteria, but many times teachers devise checklists or rating scales. Such lists or scales serve to announce criteria to the students and to allow for a more consistently applied set of standards. Many times students can use them to judge themselves or their peers. Many schools and school districts are now using such formalized rating procedures to evaluate the quality of writing and speaking of students in the school or district. In the domain of feelings and attitudes there are many commercially published interest and attitude inventories, and other measures that seek to describe students, but many teachers find these measures too general for use in a particular class. Teachers might select from among these measures and supplement them with their own questions to give the additional or more specific information that they want.

Files or Portfolios of Student Work. Keeping student work or samples of it drawn from various critical points in the school year (or over the school years) can help a teacher form a clear picture of the progress of a student. Files or portfolios may well be the best means available for checking the validity of test scores of students. If a teacher can present tangible evidence of student growth to

place against a lack of gain on test scores, such evidence should—to all reasonable persons—cast doubt on the accuracy of the test results, or point to the specific limitations of the tests themselves.

Interviews. Particularly as individualization of instruction increases, the teacher needs to talk with individual students or small groups of students to ascertain their needs, to determine progress, and to evaluate instruction. These interviews require careful advance thought and planning on the part of the teacher; they should also be flexible enough in format to be shaped, in part, by the responses of the students. A teacher might ask students what they think they have learned and then ask more specific questions about whether they have learned what was intended.

Observation of Classroom Behavior. In diagnosing student needs or checking the ongoing success of instruction, teachers can use their own eyes and ears. For example, the act of walking around the room and watching students begin a writing assignment can often enable a teacher to see which students do not understand the nature of the task set for them. The teacher can then help these students perform better and can evaluate student progress. Because of the large number of students teachers must work with and the long time span they work together, classroom observation can be made more valuable for evaluation purposes through informal and formal modes of recording data. For example, a teacher selects some particular overt action that indicates achievement of a goal and then keeps a checklist to record whether and when each student performs the action. A checklist of attitudes towards reading might include these items: visits the library voluntarily, asks for a book to read, asks for another book by an author read in class. A checklist could also be made for the demonstration of skill or understanding. In addition, much valuable information can be gained through chance encounters between students and teachers if the teacher is receptive to these opportunities both in and out of class.

Games and Contests. Many of the activities in school can be viewed as games. The winners of those games are students who have succeeded. Winning a writing prize is a clear indication of individual success (although the winner has probably benefitted from years of education, not one course). Within the class, certain of the newer learning games become themselves instruments of evaluation as well as of instruction. Succeeding at the game at times can indicate mastery of the content or skill for which the game is designed to provide practice. These measures of success are also often motivating and challenging. Unfortunately, however, when there are winners there are also losers; consistent losing can be disheartening. Experiences that bring various kinds of success need to be designed for all students if they are to improve their sense of themselves.

Published Tests. This large category, the concern of most of the rest of *Common Sense and Testing in English,* includes tests produced by publishers of educational materials to accompany those materials and tests produced to measure student achievement in general. Although most of these tests are "standardized tests" in that they are intended for use under similar conditions by

many groups of students, many are intended for much less rigorous conditions. Most older tests are "norm-referenced" so that a student's score is reported not as a simple matter of number or percentage of questions answered satisfactorily, but in comparison to the average score of a large number of people who took the test with the student or took it at some earlier time. Many new tests are being sold as "criterion-referenced" or "domain-referenced" tests; for these tests the score is reported as the number or percentage of questions answered satisfactorily. These two approaches to testing and reporting scores are discussed at greater length in later sections.

The Domination of Published Tests

All of these means of evaluating students' needs, progress, and achievement are worth using, and worth using well. One possible reason that tests externally produced and standardized have come to play such a prominent part in the lives of children and teachers is the failure of a number of teachers to find the time to devise and use their own means of evaluation well. Another reason comes, as has been said, from the demand for information, and the assumption that an instrument which is accompanied by elaborate statistical data is more to be trusted than one that lacks these data. Many of those who have encouraged accountability have said that classroom teachers are unclear about what to measure and how to measure it. These critics have been unwilling to accept the judgment of English teachers, calling it "subjective." There is some legitimacy in that charge, although the judgment of English teachers is no more subjective than the judgment of many others such as food or livestock judges at a county fair. English teachers can evaluate their students clearly and in language understandable outside the classroom. The means, variety, and quality of this evaluation are limited only by the understanding and creativity of teachers and the conditions under which they work.

Two steps are now necessary. The first is for teachers of English to make clear to administrators and the general public the failure of standardized and other published tests to reflect adequately the achievement of their students. The next few pages can help in this respect. The second step is to provide clear alternatives. The preceding section has given some indication of possible alternatives. Which alternative might be used and what the criteria of achievement might be are the province of school administrators in consultation with individual teachers. They must determine these criteria in the light of their goals. If this can be done in a professionally responsible manner so as to convince the larger public that the alternatives are in many cases better than standardized scores, teachers will have passed the test.

STANDARDIZED TESTS AND THE MEASUREMENT OF ENGLISH INSTRUCTION

Norm-Referenced Tests

Most standardized tests are *norm-referenced tests;* they report a student's score by comparing that student with other students. The most fruitful uses are to predict future achievement of individuals and groups and to examine the relative performance of large groups of students in order to locate potential problem areas for those groups. Since the tests are designed to emphasize differences among people, extreme caution should be exercised to avoid the unfair comparison of class with class or school with school within a district on the basis of these tests. After the tests have been used to locate potential problem areas, other tests are needed to determine the nature of the problem.

There is also another consideration. The norm cannot be considered a "fair" basis for comparison of groups in any case unless the tested population is very similar in all respects to the normative population—a condition which rarely exists. If the tested population differs in motivation, experience, background, or in any other important way from the normative population, using the norms stated for the test as goals to be attained becomes inappropriate. Norms exist to permit comparisons and for no other purpose, and they often do not permit fair comparison.

Norm-referenced tests are designed to be general and to show differences. They present two major difficulties to those who wish to use them to assess the effectiveness of an instructional program. First, the content of most of these tests is frequently not well defined. Because test publishers wish to appeal to a national market that contains diverse local approaches to the English curriculum, they hesitate to define content sharply. States or local districts are more likely to adopt tests if they can assume that the content of the tests matches the local curriculum. Second, construction of a norm-referenced test demands that student scores be variable so as to relate one student's score to an average, to put that score in a curve* of scores. Therefore, question or item selection is not likely to be based on program priorities, but rather on which questions discriminate among

Central Tendency and the Normal Curve: Ways of describing how an individual score relates to the average. On many tests a single score means little unless that score is related to the whole range of scores that people who took the test earned, as well as to where most of the scores are bunched. Standardized norm-referenced tests are designed in a fashion such that most of the scores achieved will be close to a middle and fewer and fewer will be close to either extreme end of the range. This phenomenon is graphically described by what is called the *Normal Curve* or the *Bell-shaped curve.* The scores are reported in terms of the *central tendency* or the averages: either the *mode,* the score that occurred most frequently; the *median,* the score that exactly separated the top half from the bottom half; or the *mean,* the arithmetic average of all the scores. On the normal curve, the mean is depicted by a line that goes to the highest point on the curve. The area of the curve is arbitrarily divided into six segments which are used to show

students. Questions might be written by English specialists, but frequently test-makers will reject or revise them in order to make the test achieve variability. Questions that discriminate among large numbers of students tend to be questions of general knowledge or general intelligence. As a result, the test may no longer represent the heart of the subject area and can no longer demonstrate the results of instruction.

Most norm-referenced tests are designed for group use. They provide (or did provide when the test was normed) statistically accurate measures of group achievement. Because they are designed primarily for groups and because of the standard error of measurement*, however, these tests are far less accurate in determining any individual's absolute level of achievement. It is, therefore, dangerous and at times grossly unfair to assign students to a category or class on the basis of a single score from such a test. This unfairness is particularly true of such tests as college admissions tests which have a high standard error.

Norms should not be confused with goals. Test users sometimes speak of bringing a student or group of students "up to the norm," thus equating a norm with a goal of instruction. To understand the fallacy and the danger in this approach, it is necessary to know something about how norms are derived. When a test has been prepared, it is given to a large group of students, usually numbering in the thousands, and the raw scores achieved by this group serve as data for preparing the norms. For any given subgroup of these students (e.g., the ten-year-olds, the fifth graders) their median score becomes the norm. Thus it is clear that the norm represents what, on the average, these children *are* achieving, not what they *could* achieve given ideal (or even improved) conditions. In one standardized test of knowledge of literary works and authors, the norming

---

the diversity or *variability* of the scores. These segments are mathematically determined and are referred to as *standard deviations* or the average distance that scores are away from the mean. Standard deviations are so computed that whenever one adds the number of scores one standard deviation above the mean and the number of scores one standard deviation below the mean, one will have included about two-thirds of all the scores on the test. Both the curve itself and the standard deviation are products of mathematical reason, not of actual results. They are useful for other kinds of statistical analysis, and often the scores are arranged so that they will fit a curve, and tests are often designed to insure that scores fit the curve. There is no proof that learning—or even intelligence—fits this curve.

**Reliability and Error of Measurement:* Reliability is the degree to which a test is an accurate gauge of an individual's performance. A person taking the same test twice might not get the same score for a variety of reasons, hence the question: "How close is the score to that person's *true score*?" An index of reliability ranging from 0.00 (unreliable) to 1.00 (absolutely reliable) can be determined, and from that a *standard error of measurement,* an index of the range in which a particular score might lie can also be determined. On some I.Q. tests, for example, the standard error of measurement is 5, which means that a student whose score is 100 might easily have gotten 95 to 105 or anywhere in between. The standard error of test scores (which should be in the test manual) must always be known and understood by any user of test information.

population averaged 30% correct. It was obvious they had not been taught the material on the test; yet their poor score became the norm and the "goal" of instruction. Thus to regard the norm as a goal may be to aim for mediocrity.

What Aspects of English Can Be Measured with Standardized Tests?

Since standardized tests are machine-scored, usually norm-referenced, and usually general measures of ability or achievement, their use is limited to those areas for which such types of measurement are valid*.

With varying degrees of success standardized machine-scored instruments, usually norm-referenced, have been developed for measuring many aspects of progress in English. Those aspects include:

Reading Skills. Decoding and word analysis; word meaning; literal comprehension; simpler kinds of interpretation and inference.

Understanding of Literature. Knowledge of facts about literature (authors, plots, literary types and devices); simpler kinds of analysis of literature.

Grammar and Language Facts. Knowledge of these facts; spelling; punctuation and capitalization.

Editing in Writing. Editing skills, particularly with respect to standard usage.

Listening Skills. Short term memory, comprehension.

Study Skills. Use of a dictionary, map reading, library skills.

Many of these aspects of English can be defined operationally in a manner generally acceptable to teachers. It is becoming less and less possible, however, to achieve common acceptance of what constitutes standard usage, the analysis of literature, and knowledge of the facts about literature. Particularly as various cultural minorities have gained their long-deserved recognition by schools and curriculum-makers, common acceptance of a standard language and culture has been challenged. Tests tend to place matters in categories of right and wrong rather than in categories of appropriateness to a group or to a situation. The failure of tests to reflect the diversity of our society has led to discimination against

*_Validity:_ The degree to which a test measures what it is supposed to measure. Validity can be defined in terms of content (does a test deal with what is taught and in a manner similar to the way in which it is taught?). Many a test called an English Test asks questions about certain matters of etiquette in formal English (like the split infinitive) that a number of teachers no longer teach or consider important. That test might be a valid measure of language etiquette but is not a valid test of what is taught. Similarly, if a test uses terms from traditional grammar and students are studying generative grammar, then that test is not a valid measure of the students' knowledge of grammar. Another way of determining validity is to establish some criterion and then determine the relation of the test to that criterion. If scores of students on a multiple-choice test of literary analysis fall in the same order as do their grades on essays analyzing literary selections or their grades in a literature class, then the test can be called a valid indicator of achievement in literary analysis.

minorities in school placement and classroom grouping. This discrimination, in turn, has increased the disaffection of minorities with school procedures and curriculum.

How well do the tests measure these aspects? Although these are the main aspects of English measured in standardized tests, one should not assume that these aspects are tested adequately. In many cases, particularly in reading tests, the pieces to be read stand in isolation rather than in some larger context, such as a story or an article, so that the use of that larger context is not tested. The same is true of tests of editing, which often contain sentences in isolation, so that a student must infer the stylistic context within which to consider the aptness of each sentence. Another problem with these tests is that they often deal only with the simpler and less subtle aspects of the process being tested, such as surface comprehension and relatively simple inferences. The standardized machine-scorable format does not lend itself easily to testing subtlety of meaning.

Limitations

Inevitably, there are some areas standardized tests don't measure. If it is true that standardized tests are to play a major role in decision-making in the field of English, it is important to note the very serious weakness in the quality of tests (or the lack of such tests) for measuring such areas as the ability to:

- appreciate literature
- organize ideas and compose in speech or writing
- vary the use of language to express thought and feeling
- speak clearly and effectively
- listen for meaning
- produce in various media
- understand media
- read critically
- deal with variations of usage and/or dialects
- express values

These limitations result in some bad effects as well. Teachers have little incentive to strive to be creative and exciting in their teaching if they know that their students will be judged solely in terms of machine-scored tests which deal with the simpler aspects of their subjects. Students have even less incentive to be creative in their uses of language, to listen to new ideas, and to engage in the lively process of learning if they know that such matters are not measured by the tests which determine so much of their future.

Measurement should reflect the best practice in teaching and the best aspects of the curriculum. Teachers can assess their students' composing abilities, their ability to deal with ideas, their organizational skill, and their creative endeavors. Although English teachers should accept those standardized tests that validly measure what they are teaching, they should insist that these tests cannot measure all aspects of students' wide range of experience in English, and that they signally fail to measure many of those aspects deemed most important by teachers, parents, and employers.

Although the demand for schools and teachers to be accountable is reasonable, teachers and administrators must resist the use of standardized test scores alone as indices of accountability. Using these alone limits the range of English instruction and places emphasis on aspects of English that are relatively unimportant. Teachers must continue to insist that there are some judgments that only they can make and some that only the students themselves can make—perhaps many years after their school experience is over. These judgments are valid and must be publicly recognized as such.

CRITERION-REFERENCED AND DOMAIN-REFERENCED TESTS AND THE MEASUREMENT OF ENGLISH INSTRUCTION

Two recent phenomena in the field of testing are "criterion-referenced" and "domain-referenced" tests. Both kinds of tests are based on a principle quite different from that of "standardized" or norm-referenced tests. Norm-referenced tests focus on the differences among scores achieved by people and thus on the differences among people. Criterion-referenced and domain-referenced tests focus on people's degree of mastery of subject matter or on people's ability to perform skills. Since comparison of students is unimportant, these tests do not have to be standardized and do not have to refer to large groups of students who have taken the test.

Criterion-Referenced Tests

Criterion-referenced tests start with an objective that someone decides students should reach. The objective is stated in quite specific terms, such as being able to recognize metaphors in poetry. A test may be created which contains fifteen passages with metaphors and five passages with no metaphors. Students are asked to sort passages, and are scored on the number they sort correctly. A successful student might be one who gets 80% correct. For a group of students, scores would be reported in terms of the percentage of students succeeding on the test at the 80% level. The 80% correct might be the *criterion,* and the report might say that 90% of the students reached the criterion. No further reporting is done, except to say who or how many reached the criterion.

A criterion-referenced test can be an effective measure of a student's progress toward specific goals. It can also help indicate the effectiveness of a given program, because it alerts teachers and other evaluators to those goals that students fail to achieve. Criterion-referenced testing has been related to highly sequenced programs and is frequently used for monitoring an instructional program as it goes along.

The limitations of commercial criterion-referenced tests are several. The first is that the objectives measured by the tests may not fit the objectives of instruction in a school. A second limitation is that criterion-referenced tests divide the world of English into tiny fragments of learning. Strictly speaking, the test on metaphor should be matched by separate tests on simile, personification, metonymy, and other devices of figurative language. The test also does not deal with understanding metaphoric language. Several complementary tests might be needed and the students would then be over-tested. While norm-referenced tests might define English vaguely, criterion-referenced tests may go too far the other way.

Domain-Referenced Tests

Domain-referenced tests were created to strike a balance between fragmentation and fuzziness. A test-maker defines a domain of learning, and criteria for success within that domain. A domain-referenced test in literature, for example, could deal with the ability to recognize and discriminate among the common

types of figurative language rather than to recognize metaphors. The test provides sufficient opportunities for the student to demonstrate mastery of the domain. It might ask students to indicate metaphors, similes, personifications, and metonymies in a variety of passages of prose and poetry. The score is reported in much the same way as in criterion-referenced tests: a certain number of students demonstrate mastery of the domain by getting a certain percentage right.

Let us suppose that a secondary school has a unit on composition which has as its main focus "point of view." The unit might deal with identifying the point of view in a variety of types of writing and with writing paragraphs from different points of view, including fixed points, moving points, and multiple points. A published domain-referenced test on point of view might have a variety of passages and ask students to identify the point of view from which each was written. It might set as a level of mastery 80% correct responses on this test. The school could use the test as a part of its evaluation program, since it measures a part of the objectives of the unit. The test does not measure writing, however, and should not be accepted as the only tool for evaluation.

Cautions in Selecting Criterion-Referenced or Domain-Referenced Tests

The primary concern of criterion-referenced or domain-referenced testing must be with what is actually taught. Although the test may be prepared by a test publisher, if that publisher defines the domain, the limits of content, and the criteria for mastery, a school whose curriculum has a similar domain similarly defined may well use the test for evaluation of both students and curriculum. Teachers in a school should examine any domain-referenced test extremely carefully before recommending that the school use it. If students have not had an opportunity to learn what is tested, test results will not say anything more than that the students did not learn what they were not taught.

Both criterion-referenced and domain-referenced tests are relatively new. Their abuses have not yet emerged. Among the potential abuses is the one that the criterion-referenced test will turn into a norm-referenced test by having the results reported in terms of a relationship between the score achieved by one student and the scores achieved by some "national" or regional sample of students. Teachers should beware of this misuse of criterion-referenced and domain-referenced tests; results should be reported only in terms of number of items one student did well on or number of students who reached a given level of proficiency.

CRITERIA FOR SELECTING STANDARDIZED TESTS IN ENGLISH

From recent critical reactions of teachers to standardized achievement tests currently on the market comes a clear challenge to the profession to provide leadership in the development and use of valid and reliable tests in English. The fact that many current standardized tests are not acceptable to teachers of English shows the urgency of this challenge. And simple acceptance of new conceptions of testing—criterion or domain-referenced tests replacing norm-referenced tests—could lead the profession into new testing problems rather than to the solution of more basic problems of evaluation.

English teachers have a major responsibility to insure the best practices for measuring student growth and evaluating instruction. A part of that responsibility lies in teachers defining those parts of English that are of concern to them. As many activities that go on in English classes demonstrate, English teaching contains parts that are relatively easy to define. Some of those parts, like spelling competence, are also relatively easy to measure; some, like lifelong reading habits, are relatively difficult to measure within the confines of school. Other parts of English resist precise definition, particularly those in which the imaginations of students and teachers are given free play and those which deal with the personal uses of language. Teachers can say that they want students to develop a personal style of writing; they have a hard time giving a number of uniformly and objectively verifiable instances of such writing. Developing a test or a set of criteria to measure progress in such an area is exceedingly difficult and may, in fact, be impossible.

Rather than administering a test that has been handed out sight unseen, teachers of English should insist upon examining the test to determine whether it is appropriate to their program and their students. Most tests are available for inspection, although a few, like those for admissions programs are held secret. When examining the test and the test manual, teachers should consider the issues set forth below. A checklist on pages 29-30 provides a summary of those considerations.

General Criteria for Choosing a Test

Within the range from the mechanical to the creative, there are areas of English instruction that can be defined, that have clear-cut criteria for student mastery, and that can be measured through some kind of objective test. To select or make such a test, teachers must see that three conditions are fulfilled:

1. the limits of the content can be clearly defined;
2. a means can be devised by which students can demonstrate mastery;
3. it can be clearly determined whether the student's answer is correct or acceptable.

To take an example from composition: students can demonstrate mastery of complex syntactic structures by producing combined and embedded sentences with few errors. The types of combinations and embeddings could be defined. A test could be made that provides students with a series of related thoughts, each

stated in a simple sentence, and asks the students to rewrite them as one sentence. There could be a teacher's guide that indicates acceptable responses to each problem, and the grounds for judging those acceptable and not others. If teachers cannot agree to those grounds they should not use the test.

The Right Test for the Right Use. No matter what kind of objective instrument is used, the first consideration in the selection or construction of the best test is the use that is to be made of it. Measurement of skill requires an instrument different from that required for measurement of factual knowledge or of attitudes and values; and measurement of progress toward or achievement of particular program objectives requires a different test from one used to measure more general educational achievement. Different tests should also be selected or created for diagnosing student needs, for placing the student in a program, for measuring student achievement, and for evaluating instructional programs.

Tests and test questions should try to reflect cultural and human diversity. It is particularly unfair to make judgments on the basis of a test that would discriminate against a student because that student uses a dialect different from that of the test or comes from a culture* whose values, understandings, and perceptions are different from those expected on a test. Many English tests have content totally alien to large groups of young people, such as reading passages about life in a Maine fishing town or about a world of butlers and housemaids on a test used in a big southwestern city with a large Latino and black student body. Although the language of the test is standard English, all items must be written in a style that does not inadvertently "trip up" users of those other dialects which have a consistent but different grammatical base, dialects that make up the rich mosaic of American English.

Test writers must also be sensitive to the portrayal of women's and men's roles. Widespread changes in attitudes and goals are strongly affecting young people today, making some types of test questions unacceptable.

Some of these tests are as much as 35 years old. They were unfair when they were written; they are grossly unfair today. To achieve these goals of responsiveness to human diversity and clarity of language, test-makers should seek a review of test items by concerned groups within our society and make a trial use of the test with a variety of student populations. A report on results of such a review should accompany the test. No one test publisher at this time has the resources within its editorial staff to provide the necessary sensitivity to all these

---

Culture-free and Culture-fair Tests: Terms used to describe tests that do not discriminate against persons from different cultures in a society. In English testing, where "standard English" has come to be defined as English used by certain speakers in the dominant culture, a culture-free test would be one on which those speakers would not have an advantage. This is probably an impossibility; the term *culture-fair* has been substituted to describe a test which draws questions related to different cultures and tries to place their members on an equal footing. It is quite probable that in English, a truly culture-fair test has yet to be developed.

concerns, and no individual test-writer can be aware of all the language diversity that could influence test results.

Test questions should be clear. The preparation and selection of the test questions or items* themselves, of course, are crucial. The mode of the student response (true/false, multiple choice, or some other mode) should enable the student to demonstrate achievement in a way acceptable to the profession. Two response modes may not call for the same kind of skill: if a student should create an answer to a question, the fill-in is appropriate; if the student should reach the answer by reasoning, the multiple-choice question could be suitable.

Further, the individual items themselves must be constructed with a high degree of sophistication. In true-false questions, for example, there is a danger that the answer might lie somewhere in between "true" and "false."

> "T.F. *Now is the time* is a complete statement." Grammatically, the expression is complete, but many people would expect there to be some complement to the expression, which in its present form seems to leave the reader in mid-air.

In multiple-choice items the distractors are as important as the responses keyed as correct.

> Which of the following is regarded as the "great searcher of the human heart?"
>
> a. Shakespeare
> b. Wordsworth
> c. James Baldwin
> d. Rachel Carson

Although the quotation may well have originally referred to Shakespeare, it is not easy to dismiss Wordsworth's claim, and some might even make a case for James Baldwin or Rachel Carson.

Test questions should measure what is intended. The test item must be true to the instructional objective and not bent to a shape required for statistical reasons. If a test appears too easy, test-makers often make it more difficult by complicating the wording of simple questions. They use such devices as negatively phrased questions ("Which of the following is NOT . . . ?") to trick the unwary. Items should call upon the knowledge or skill being measured and not require unrelated information or skills. The negatively phrased question often becomes a test of logic. Another example is the question in a reading test that can be answered only if the student has outside knowledge. In a reading passage about dinosaurs a question that asks how long ago they lived would be unfair unless the passage contains that information.

**Test Item:* The problem or question set for the student. The common forms of test items are true-false, multiple-choice, matching, and fill-in or completion. In many of these forms, the actual problem is called the *stem* and the correct answer or choice is called the *key.* If there are wrong answers given, as in the multiple-choice item, they are called *distractors.*

Many other factors also affect the usefulness of an item. In norm-referenced tests of reading achievement, the intent is to measure a range of reading ability, and the items must therefore represent a range of reading difficulty. In tests of any aspects of English other than reading every effort should be made to create easily readable items so that they will measure what they are intended to measure. Tests should avoid the use of jargon or technical terms with which the students might be unfamiliar. Students must be able to understand clearly the meaning or intent of the items. The item itself must be accurate and not inadvertently state false or misleading ideas.

Test questions should pose meaningful tasks. Another crucial aspect in the development or selection of items for a standardized test is student interest. Most tests are based on the assumption that students will try their best. The truth of this assumption rests on at least two factors inherent in the test items: (1) that the student sees the test experience as worthwhile, and (2) that the ways in which language is used are of functional concern to the student. If, in tests of skills, items present tasks that the student has never performed or likely never will perform or topics in which the student has no intrinsic interest, the results will probably not reflect that student's capability. A test of editing skill should deal with writing from a current good writer rather than from Thomas Carlyle, if it is to present a meaningful task to most students today.

The test as a whole should make sense. The construction of valid and reliable tests depends upon the shape of the test as a whole as well as upon the individual test items. The entire test should be a functional whole in which major aspects of the subject matter are systematically measured. A spelling test, for example, might be based on major and minor spelling patterns, or on commonly misspelled words, or on some equally acceptable principle, if the results are going to be used to generalize about a student's spelling ability. A good test of analyzing literature might have the items arranged to follow the order of the passage, not in random order, and it might have questions about specific parts of the passage preceding general questions. In a test of writing, test items might best be placed in a situation context so that the student could make judgments knowing who were the assumed writer and audience, and what was the assumed purpose.

Directions should be clear. Test directions as well as test items should be readable and clear to students. Unless they understand exactly the tasks the test sets for them, students will not be able to demonstrate their competence. Just as all tests should not be tests of general intelligence or of test-taking ability, all tests should not be tests of reading ability if true achievement in other areas of English is to be accurately measured.

The whole test should be clearly defined. Not only should all of the items on the test be faithful to the central concerns of English, but the scope of the particular test should be clearly defined. This scope should be reflected in the test title and manual; for example, an instrument entitled "English test" might better be called a test of usage, punctuation, and spelling if its content is limited to these

areas. The user of the test must be aware of which goals and objectives of the English program are measured by the test and which are not. The results then will not be used to describe progress in the totality of English, but will describe progress in an accurately defined portion of the English program.

The test manual should clearly state what the test is measuring. It ought to communicate to teachers and administrators the objectives, the content, and the usefulness of a test. Then these educators can decide whether to use the instrument for diagnosis or for periodic assessment of school or district achievement. *Their decision must be based on what the test measures.* Because no test measures all parts of the content of English, publishers should state clearly the objectives and the content of English instruction that a given test sets out to cover. With this information, test users can identify for their various audiences of accountability the part of the English curriculum for which the results of the test offer assessment information. The clear statement of objectives and of test content can assure teachers, students, administrators, and others that the test has been chosen because its objectives are also the objectives of a particular school curriculum. Thus test-givers and test-takers may place confidence in the integrity of a test program in which assessment is directed toward the outcomes of the particular learning experience.

Test manuals should report on the validity of the test. Reports of validity studies of the test should be part of the manual. The criterion against which the test has been validated should be clearly stated. If, for example, a test seeks to measure some aspect of writing ability, then the test might well be validated against information derived from writing samples from some clearly defined group. A test of mechanics of writing might be validated against a count of the errors appearing most frequently in a large number of compositions from a clearly defined group. If the measure is predictive of later performance, then the criterion should be clearly spelled out and the extent to which the test predicts the criterion should be stated.

The manual should give full information on reliabiltiy. Data on reliability for the whole test and the method by which it was established should be included. If the test to be administered is a shortened form of a longer test, the reliability coefficient for the shortened form should be given along with a clear explanation of its significance in interpretation of test results. If the test consists of subtests, subtest reliabilities should be included. If the reliabilities are high enough for group diagnosis but are not high enough for individual diagnosis or measurement, the manual should so inform teachers and administrators. Many achievement tests with attractive subtest headings really furnish no reliable information about these subareas of achievement.

The manual should give the standard error of measurement. It should state clearly the meaning of this estimate of error in relation to an individual score. Error of measurement indicates that no individual score is absolute; the standard error indicates the band within which an individual score is likely to fall if the receiver of the test were to take the test again.

The manual should give full information on norming. The manual for a norm-referenced test should also furnish complete data on norming procedures. No test can be absolutely valid or reliable. No test can truthfully claim to have national norms that represent all groups at any given grade level in the school population of the United States. But test-makers generally do try out tests with a vareity of groups of students. The test manual should furnish full information on the dates of norming and the nature of the groups of students to whom the test has been given. The nature of the groups would include such factors as section of the country, area lived in (whether rural, urban, inner-city, or small town), family income, race, ethnicity if it is a factor, and bilinguality. This information enables decision-makers to grapple with the problem of how closely the groups in the norming population match the student population whose achievement they wish to measure.

The manual should give other kinds of scores besides grade level scores. Teachers and students like to chart evidence of growth and progress. Often they use test scores as part of the evidence. If a test manual provides grade-equivalent scores, it should clearly indicate that they are poor indicators of change or growth. Although also unsatisfactory, stanines or percentile ranks are slightly better methods of estimating change or growth. Grade equivalent scores on a test normed on a population of seventh and eighth graders may reveal something about median achievement, above median achievement, and below median achievement. Though it may attempt to do so, the test cannot reveal accurately, for example, that a low-achieving seventh grader is reading at fourth grade level or that a high-achieving seventh grader is reading at twelfth grade level. Extremes of grade equivalent, such as 4.0 when the test is normed on seventh and eighth grade students, are likely to be highly unreliable. Out-of-level testing* for those students below grade level is also an unprofitable way to get assessment data. Eighth graders reading at the fourth grade level are not likely to be assessed accurately by a test normed on fourth graders. The conclusion that lower achieving students are falling further behind at successive grade levels results from using grade equivalent scores and does not necessarily reflect actual lower achievement. The reason for this is that the rate of growth for one group may be faster than for another; both are growing, but grade equivalent scores would not so indicate. Although the use of grade equivalent scores appears a simple index of achievement, it is a grossly deceptive index, particularly for higher and lower achieving students.

The manual should clearly caution about comparing scores on two forms of a test. It should define and describe in nontechnical, clear language the fallacies

---

**Out-of-Level Testing:* The use of a test created for students at one level with students at another level. If, for example, a class is given a test designed for that grade and some students score well above the grade norm, to test those students with a test designed for students two grades above would be out-of-level testing. The subject matter of out-of-level tests is likely to be inappropriate despite the ability of the students.

and pitfalls involved in comparing the results of (equating*) two forms of a test. These statistical processes seem to make possible and even reasonable the comparison of scores for different test forms. Although teachers might want to compare students' achievement at the beginning and end of the year or to compare freshman and senior achievement, accurate comparisons using different forms of a test are often not possible.

The manual should indicate the pitfalls of variability. The half century during which standardized tests have been used in this country has been a period in which variability—both in total scores, and in discriminating* power of each item in a test—has been prized as a way of identifying differences among learners. Thus the results of achievement tests offer the same bell-shaped curve representing variability as intelligence tests do. Most educators have assumed until the last decade that the distribution of student test scores, whether the scores represent aptitude or achievement, all fall into the normal bell-shaped curve. Recently, however, serious-minded educators have shown that achievement scores need not and should not be thought of as following that simple a pattern.

**Equating:* The statistical process of relating scores on two different forms of a test. If there are, for example, two forms of an eighth grade reading test, equating normally relates the two tests through questions that appear on both tests or by giving both groups of students a common test. A person is thereby able to convert the score on the two forms to one scale. If there is an eighth grade form of a test and a tenth grade form of the same test, the two forms would also have some common questions so that one could relate the scores on the two tests and place the students on a single continuous scale. The assumption of equating is that the test items and problems on two forms of tests are similar enough so that if differences in difficulty are statistically eliminated it does not matter which test a student takes. That assumption has come to be challenged by experts in testing.

**Discrimination:* The extent to which the answer to any one item predicts the score on the whole test. An index of discrimination is the mark of relationship between the right answer on an item and achievement on the test as a whole. This index also serves to point to the difficulty of the item and to the reliability of the test.

CRITERIA FOR THE INTERPRETATION AND USE OF TEST RESULTS

The careful development and selection of reliable and valid tests in English is vital to the profession; even more crucial, at least in the short run, are the informed interpretation and use of test results, and the continued re-education of the users of test results to prevent false interpretations and misuses.

Tests Only Sample Learning. Teachers, administrators, and other decision-makers for the schools must keep in mind that they represent only a limited sample of what a person does or can do, a sample drawn at a particular point in time, and a sample from which reasonable inferences can be made. For example, we can infer that a student who chooses correct answers to items on a valid test that is designed to measure analytical skills in literature is likely to be able to read literary selections with some degree of discernment at some time close to when the student takes the test. We can make no assumptions about the interest or taste in literature of that student or about the breadth of the student's knowledge of literature.

Avoid the use of test scores and nothing else. Test results in isolation should not be used for decision-making. There must be clear recognition of the relationship of the test score to other information about the student, such as the student's opportunity to learn what is tested, or the fact that the student is a non-native speaker of English. Test results provide but one set of information, information which is useful and usable to the extent that the test is valid and appropriate for a given purpose. At present, because testing has a kind of mystique, particularly in the statistics and terminology which surround it, teachers often need the help of a qualified specialist in interpreting test results. However, the more knowledgeable classroom teachers themselves become about test construction and interpretation, the better for the appropriate use of test results. Informed teachers should participate in the interpretation of test results as well as in the construction and selection of tests. Informed teachers have the right to test the tests, rejecting the invalid and inappropriate ones and guarding against zealous misinterpretation of the valid and appropriate ones.

Tests should be given only for purposes established beforehand. Because tests in their infinite variety are available and testing has become one of the expected activities in schools, it is important that tests not be given from habit and under the assumption that testing is a good and expected activity. There is even a danger that a teacher can accumulate more test data than can possibly be used. School systems should, whenever possible, adopt patterns for a testing program that will provide information and eliminate waste, both human and financial. For example, all tenth graders might take a brief screening test in reading skills; those falling below the 75th percentile might take a more in-depth test; those falling below the 40th percentile in this second test might be given a series of diagnostic tests. These tests form a starting point for instruction in an area of English the teachers consider important. That is their only use at that time.

Use the right test for the right purpose. Uses of test results should, in part, depend upon the nature of the test. In general, standardized achievement tests are most appropriately used in schools to compare groups, not individuals. Informal measures and diagnostic tests are best used to locate learning difficulties. Criterion-referenced or domain-referenced tests are best used to determine individual mastery and growth or group progress towards a specified goal. Scores should be used only for the purpose for which the test was designed.

Growth should not be measured by averages alone. If standardized tests must be used to measure change or growth, a function which, as has been stated, they do not serve well, users should make use of performance information other than mean or median scores. For example, if 25% of the students initially fall in the lowest stanine but only 10% score in that range after instruction, a significant improvement may have taken place even though the mean for the whole group stayed close to what it was at the beginning. Reporting the mean alone would show no improvement and falsify the view of the students.

Test scores should not be used out of context. Test results should not be used, even in a limited way, to compare schools or school districts without weighing in such factors as location (rural, urban, or suburban) and the socio-economic status of the population involved. Weighing in these factors is not a means of providing excuses for poor performance, but does help one understand the relationship of local scores to national norms.

Test scores should not be labels. They should not be used for the labeling of students or as the sole criterion for the placement of students in special sections. The results of one test are not reliable enough to serve as the basis for such important and far-reaching decisions. Because one test is not enough and because a test result represents only a snapshot of a student on one day, placement decisions should be reviewed at least once a year for possible changes in assignments.

Test score results do not evaluate teacher effectiveness. As has been pointed out repeatedly, achievement test scores represent only a fraction of the effect of English teaching and learning. Thus they measure a fraction of an English teacher's goals. Because of this major limitation, achievement test scores of students should not be used to evaluate individual teachers. They are too unreliable an index to be used for personnel decisions.

Test norms may not match local conditions. Test publishers and test administrators have a special responsibility not only for making sound tests but also for attempting to ensure the appropriate use of their instruments. For example, on test score reports clear warning should be given against the use of the scores apart from other information about the students. Since publishers' norms, erroneously called "national norms," are of limited value to teachers or local school systems, strong consideration should be given to developing more specific norms, frequently updated, for use in interpreting test results for regional and other sub-populations (e.g., northern inner-city students, or southern rural students.) Further, districts should create their own norms.

Test reports should not confuse norms with goals. The fallacy of "bringing a student or group of students up to the norm," thus equating a norm with a goal of instruction, has been described. Those who report test scores have an obligation to point out this fallacy clearly.

Test results should be published with adequate cautions. School officials should not release to the media test score results without a clear statement of what they represent. If scores based on norms are reported, the press release should include information similar to that in the following paragraph.

> These scores are reported in comparison with a group of (number) children who were tested on (the same) (a different) form of the test in (year) . The children came from (locations) . Any comparison with that group must consider differences in school and community conditions and curriculum. An individual student's score might actually be ___ points above or below the score reported for this test. The scores of the group with which the children are being compared are not to be thought of as a goal of ___ 's educational program but are used as a rough comparison only.

If the scores are based on criterion-referenced or domain-referenced tests, the paragraph should include:

> The tests on which these scores are based represent a (national) (local) panel's consensus of what students should be able to do at the end of grade ___ . The panel indicated that these tests cover (indicate portion) of the goals in the (English) (reading) (language arts) program. Giving acceptable answers to ___ % of the questions or problems on the test was judged to indicate mastery of the field. Success in other goals is indicated by _____ . Information about students can be obtained by writing or calling _____ .

Such paragraphs may take away some headlines. They may also help the media and the public to consider fully the lack of faith that must be placed in test scores taken out of context.

APPENDIX A

CHECKLIST FOR EVALUATING ENGLISH TESTS AND TEST USES

Name of Test _____

Date when test was made and latest revision _____

Content or Skill Areas of English for Which Test is Designed _____

Test Content

Do items represent the content or skill area adequately? _____

What parts of the area are omitted? _____

How important is the area measured to the curriculum of the school? _____

Does providing the correct answer require the skill or knowledge tested? _____

Is the answer format an appropriate index of the skill? _____

Is the test label accurate? _____

Test Format

Are the items clear or are they ambiguous in wording? _____

Are the items likely to be of interest to students? _____

Are the items responsive to human diversity? _____

Are the test instructions clear? _____

Does the test seem to have a logic or sequence that is appropriate? _____

Is the language of the test current? _____

Test Manual

Does the manual describe the purpose of the test? _____

Did teachers help construct the test? _____

What checks of validity have been made? _____

Are specific items related to specific objectives? _____

If test is norm-referenced:

What is norming population? _____

Does it match local conditions? _____

Is test the same form of the test used for norming? _____

What is reliability of test? _____

of subtests? _____

What is standard error of measurement? _____

How are scores reported? _____

Are percentiles or stanines given as well as grade levels? _____

Are there adequate warnings about misrepresentation of score reporting in the manual? _____

If test is criterion-referenced or domain-referenced:

Has the criterion been defined clearly? _____

Is the criterion appropriate to the curriculum? _____

Is the criterion appropriate to the student's level? _____

Do the items measure the criterion? _____

Test Reporting

Is there a clear statement of relation of test to program objectives? _____

Is there a clear statement that teachers helped choose the test? _____

Is there a clear statement of the limited inferences that can be drawn from the test? _____

If a norm-referenced test:

Is there a clear statement of relation of test scores to publishers' norms? _____

to local norms? _____

Is there a clear explanation of what scores mean? _____

of the problem in interpreting grade scores? _____

of the fact that scores provide a description of groups, not individuals? _____

APPENDIX B

CITIZEN'S EDITION: COMMON SENSE AND TESTING IN ENGLISH

Every year in school, children study English, learning to read, to write, to speak, and to listen. Children learn skills in language in order to express themselves more efficiently and to understand others better. They learn about spelling, grammar, books, authors, plays, and films. Their learning includes not merely knowledge but habits of reading or not reading. They acquire attitudes toward books and writers, toward honesty in their own writing, and toward the best use of the language that is their native or acquired heritage.

In school these children also take tests called "English Achievement Test" or "Standardized Reading Test" or something similar. These tests cover a few of the skills of reading, some knowledge of grammar and spelling, and occasional bits of information about books, authors, and libraries. Very few of the many skills they are taught are measured in these tests; much important learning is not covered. Students are often tested about obscure bits of information that they have not studied.

On the basis of tests that skim the surface of their learning, children and their parents are told that one child is two years above grade level and another is a year and a half behind. Whole schools are ranked and whole groups of children are branded with labels like "slow learner" or "underachiever."

Standardized achievement tests in English are usually treated as though they were the ultimate word on the ability and performance of our children; yet they give inadequate information to administrators, parents, teachers and children.

A student and the student's teacher and parents want to know how well that student is doing in the specific requirements of a course. Can the student read the books that are assigned? Can the student write as clearly as the other members of the class? Does the student need special help? Is the student interested in what is going on in class?

Such specific questions can best be answered by tests that the teacher makes up, by the teacher's grades on the student's work, by student self-judgment, by files of work, questionnaires, interviews, and observation.

A school board or a state school office, or the federal government want much less specific information. They want to know how well most of the children at a given level like seventh grade are doing in reading and writing. So they make up or purchase tests that are designed to be given to many students—usually a multiple-choice or a true-false test.

Most standardized tests compare a student, a class, or a school to a national average. This comparison has many faults: it assumes that any group of students will always vary from very good to very poor, no matter what they have been taught or how well; and it insures that some do well and some poorly by asking questions of varying difficulty (some about things a student might not have studied). The questions deal with general topics in English rather than the specific topics that students in a particular school might have learned. It asks questions regardless of whether the material is even taught in that school. Nationwide or statewide tests cannot represent the specifics of a particular school's English program.

Big national tests are limited tools for measuring how well our children are doing. Using them to judge an individual student or to compare the students in a single class is like using the scales in a truck-weighing station to measure whether everyone in the family gets the same amount of ice cream.

Information about a Student's Achievement in English Can Come from:

Tests made up by the teacher or the department in the school.
Papers and other work that has been graded by a teacher or group of teachers.
The student's self-judgment or judgment by other students.
Questionnaires asking for opinions, interests, attitudes.
Files or portfolios of student work.
Interviews and oral examinations.
Observations of classroom performance.
Games and contests.
Published tests.

LEGITIMATE USES AND USERS OF THIS INFORMATION

RESULTS ABOUT THE . . .	MAY BE USED BY THE . . .	TO HELP . . .
the individual student	student student's parents student's teacher student's counselor	evaluate individual progress make decisions about teaching make predictions about future work
a single class	students in class teacher of class building administrator	evaluate class progress improve programs
students in a school	students in school teachers in school building administrators district administrators	evaluate programs identify needs
a school district	teachers in district administrators in district public of district state administrators	evaluate programs identify needs
all the schools in a state	district administrators state administrators federal officials general public	identify needs assess progress
all the schools or students in the nation	state administrators federal officials general public	identify needs assess progress

Some Questions to Ask about the Published Tests in English Your Children Are Taking

Do the tests ask your children to know or to do the same kinds of things they are being taught?

Do the tests discriminate against children because they do not speak a particular type of English or because they come from a particular part of the United States?

Are the questions clear? Are the questions designed to trick students rather than to test them honestly about what they have been taught?

Have the people who published the test proved that the test measures what experts in English think should be learned or that it measures knowledge of a skill in English accurately? Have they explained how trustworthy a score is?

If the test refers to a "norm" (an average group who got a particular set of scores) against which your children are to be compared in their uses of English, have the test publishers clearly identified that group so that you can tell whether your children should be compared to them?

Have the test publishers warned all those who look at their scores that "grade levels," particularly in English and reading, are not absolute standards but very rough indicators of averages? (Many professional associations have urged the abolition of grade-level reporting of test scores because of their misuses in the schools.)

Does the school indicate what action teachers should take for a child with a particular score or for a group with a particular average?

Have the test publishers and your school administrators warned against relying on test scores alone to tell you how well one child or a group of children is doing in reading or writing?

Have your school administrators cautioned about misinterpreting test scores and not using them to label students or to label teachers? Most national tests cannot be used for these purposes.

Have your school administrators cautioned the media about not making headlines over the unreliable information presented in test scores and averages?

You should bring these questions to the attention of teachers, school administrators, and citizens' groups. If they cannot answer these questions satisfactorily your children are in danger of having their lives ruined by bad tests badly used.